Can we trust PCA on nonstationary data?

Yanrong Yang

Australian National University
E-mail: yanrong.yang@anu.edu.au

Abstract: This paper investigates the asymptotic distribution of the spiked empirical eigenvalues for high dimensional complicated data, which take into account various structures of the population covariance matrix, dependent sample observations and large dimensionality. It provides new insights into three important roles that play in principal component analysis (PCA): the leading population eigenvalues, dependent sample observations and dimensionality. A surprising discovery is that spiked empirical eigenvalues will reflect the dependent sample structure instead of the population covariance under some scenarios, which indicates possibly inaccurate dimension reduction from PCA for high dimensional data. In particular, we show some modern statistical methods fail in estimating the number of spiked population eigenvalues for high dimensional data with factor model structure and dependent sample observations. To make further study, we propose a test statistic to distinguish spiked population covariance structure from dependent sample structure, especially for high dimensional time series with unit root. Our results are successfully applied to OECD health care expenditure data and US mortality data, which illustrate nonstationary strong temporal dependence. We provide justification for popular literature on mortality forecasting, in which PCA is applied on mortality data directly.